STAT 672: Homework 4

Tom Wallace

April 1, 2018

A

See `hw4.py` for code.

It is recommended to scale data when using SVM because if we do not, we essentially penalize each feature differently. To quote Hsu, Chang, and Lin (2008), "the main advantage [of scaling] is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges" [1].

This can be shown analytically. Consider a typical soft-margin SVM classification problem.

$$\min_{w_0,w} C\frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - Y_i(w_0 + \langle w, \Phi(X_i)\rangle)) + \frac{1}{2}||w||_2^2$$

Let $\lambda = \frac{1}{2C}$ and assume for convenience with no loss of generality that centering has occurred and so no intercepts are needed.

$$\min_{w} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - Y_i(\langle w, \Phi(X_i)\rangle)) + \lambda||w||_2^2$$

Consider a diagonal matrix S whose non-zero entries consist of the respective L2 norms of the columns of Φ:

$$S := \text{diag}(||\Phi_{\bullet 1}||_2 \dots ||\Phi_{\bullet D}||_2)$$

Restate the SVM using S:

$$\min_{w} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - Y_i(\langle wS, S^{-1}\Phi(X_i)\rangle)) + \lambda||w||_2^2$$

Define $\theta = wS$:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - Y_i(\langle \theta, S^{-1}\Phi(X_i)\rangle)) + \lambda||\theta S^{-1}||_2^2$$

We now can see that feature with a small numerical scale are penalized more heavily (conversely, features with a large numerical scale will have greater emphasis in the final classifier).

$$= \min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - Y_i(\langle \theta, S^{-1}\Phi(X_i)\rangle)) + \lambda\sum_{j=1}^{D}\frac{w_j^2}{||\Phi_{\bullet j}||_2^2}$$

B

See `hw4.py` and `output` for code and output.

C

The number of support vectors decreases as C increases, as depicted in Table 1.

Explanation: C is the penalty on slack variables ζ, which capture the degree to which the margin constraint is violated. A small value of C means a small penalty on margin violations, and will result in a relatively wide margin and many points laying on or inside of the margin (and hence a high number of support vectors). A large value of C means a large penalty on margin violations, and so the separating hyperplane will have a much tighter margin, meaning fewer points that lie on or in the margin and hence a smaller number of support vectors.

The above explanation also covers why run time for training increases with C, eventually resulting in non-convergence. When C is small, the separating hyperplane is easy to find since margin violations are only lightly punished, and so training run times are low. As C becomes larger, the SVM tries more and more to find a separating hyperplane that doesn't result in margin violations, and requires more calculation (run time) to find. Once C attains a sufficiently large value, the SVM may fail to converge because margin violations are unacceptable (due to the high value of C) but the data are not linearly separable and so some margin violation must be accepted, resulting in a catch-22 that the optimization routine cannot overcome.

Table 1: Influence of C on number of support vectors

C	**n(SV)**
10^{-1}	1818
$10^{-0.5}$	1777
10^{0}	1443
$10^{0.5}$	1083
10^{1}	847
$10^{1.5}$	676
10^{2}	576
$10^{2.5}$	511
10^{3}	478
$10^{3.5}$	456
10^{4}	448
$10^{4.5}$	448
10^{5}	439
$10^{5.5}$	438
10^{6}	441

D

w^* can be interpreted, but one must be cautious. A positive w means the corresponding feature is associated with the positive class, while a negative w means the feature is associated with the negative class. The size of w, within the context of a particular model, indicates the magnitude of this association (e.g., if one w is much larger in absolute value than other w, it is strongly associated with a class, and probably is very useful for prediction). However, the caveat "within a particular model" is crucial. Different regularization parameters will produce different coefficients for identical datasets. We cannot make the sort of specific interpetation that is possible with some parametric models, e.g. in linear regression "every one inch increase in height X_1 is associated with a β_1 increase in weight (lbs)."

Table 2 depicts the five largest and five smallest weights obtained with $C = 10^{4.5}$. We can interpret these weights. Emails that TALK IN ALL CAPS and mention money (U.S. dollars) are especially likely to be categorized as spam. Emails in which the sender mentions the name of someone that we presumably both know (George) and refers to meeting (presumably in-person) are indicative of a "real" relationship and hence are especially likely to be categorized as not-spam.

Table 2: Largest and smallest weights

Feature	**Weight**	**Feature**	**Weight**
3d	177.47	george	-342.13
capital_run_length_average	144.75	hp	-162.41
$	45.77	415	-161.59
857	33.01	cs	-140.11
000	30.43	meeting	-56.00

E

As shown in Table 3, relatively small values of C are associated with false positive rates about 1%. If we feel very strongly about staying below 1%, C=$10^{-0.5}$ has a FP rate of about 0.05%. If we are willing to go slightly above 1%, C=10^0 has a FP rate of about 2% and has over double the TP rate.

Table 3: True and false positive rates for selected C

C	**FP rate**	**TP rate**
10^{-1}	0.00	0.0004
$10^{-0.5}$	0.005	0.121
10^0	0.020	0.256
$10^{0.5}$	0.026	0.309
10^1	0.033	0.0336

F

We can modify the soft-margin SVM to incorporate different costs for different forms of misclassification. Our typical set-up assigns the same cost C to margin violations (e.g., misclassification) for both classes:

$$\min_{w\in\mathbb{R}^d, w_0\in\mathbb{R}, \xi\in\mathbb{R}^n} \frac{1}{2}||w||^2 + C\frac{1}{n}\sum_{i=1}^{n}\xi_i$$
$$s.t. \quad Y_i(\langle w, x_i\rangle + w_0) \geq 1 - \xi_i \quad \forall i$$
$$\xi_i \geq 0 \quad \forall i$$

We can instead choose to assign different costs C_+ and C_- for margin violations by the positive and negative classes, respectively.

$$\min_{w\in\mathbb{R}, w_0\in\mathbb{R}, \xi\in\mathbb{R}^n} \frac{1}{2}||w||^2 + C_+\frac{1}{n_+}\sum_{i|y_i=1}\xi_i + C_-\frac{1}{n_-}\sum_{i|y_i=-1}\xi_i$$

In our case, the negative class corresponds to not-spam while the positive class corresponds to spam. We feel that it is highly undesirable to miss important email. We can operationalize this by more heavily penalizing margin violations by the not-spam class.

$$C_- >> C_+$$

References

[1] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. *A Practical Guide to Support Vector Classification*. 2016.